DOCUMENT RESUME

ED 234 375 CS 007 317

AUTHOR Johnston, Peter
TITLE Prior Knowledge and Reading Comprehension Test Bias. Technical Report No. 289.
INSTITUTION Bolt, Beranek and Newman, Inc., Cambridge, Mass.; Illinois Univ., Urbana. Center for the Study of Reading.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Sep 83
CONTRACT 400-76-0116
NOTE 57p.
PUB TYPE Reports - Research/Technical (143) -- Information Analyses (070)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Elementary Secondary Education; Grade 8; *Prior Learning; *Reading Comprehension; Reading Diagnosis; *Reading Research; *Test Bias; *Testing Problems; Test Interpretation; Test Items; Test Reviews; *Test Validity

ABSTRACT

To show the difficulty of eliminating test bias and to develop a methodology for distinguishing between the effects of prior knowledge and of skill development on reading comprehension, 207 eighth grade students from rural and urban areas were administered an 18-question reading comprehension test. Quantitative and qualitative effects of prior knowledge on reading comprehension were demonstrated through an examination of student performance on the test's different types of questions: (1) textually explicit--drawing on information directly stated in a single sentence of text, (2) textually implicit--requiring a synthesis of information, and (3) scriptally implicit--demanding background knowledge. The study suggests that test scores are biased by prior knowledge and reflect the students' I.Q. more than specific reading comprehension skills. The findings indicate that test bias can be lessened by asking central, rather than peripheral, questions on passages for comprehension and by using a content-specific vocabulary test to estimate the individual's prior knowledge. (MM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED234375

CENTER FOR THE STUDY OF READING

Technical Report No. 289

PRIOR KNOWLEDGE AND READING COMPREHENSION TEST BIAS

Peter Johnston
State University of New York at Albany

September 1983

University of Illinois
at Urbana-Champaign
51 Gerty Drive
Champaign, Illinois 61820

Bolt Beranek and Newman Inc.
50 Moulton Street
Cambridge, Massachusetts 02238

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

The research reported herein was supported in part by the National Institute of Education under Contract No. HEW-NIE-C-400-76-0116 while the author was at the Center for the Study of Reading at the University of Illinois. The paper is part of the author's doctoral dissertation and considerable thanks are due to (alphabetically) Dick Anderson, Bob Linn, George McConkie, David Pearson, and Peter Winograd.

CS 007317

EDITORIAL BOARD

William Nagy
Editor

Harry Blanchard
Wayne Blizzard
Nancy Bryant
Pat Chrosniak
Avon Crismore
Linda Fielding
Dan Foertsch
Meg Gallagher
Beth Gudbrandsen
Anne Hay
Patricia Herman
Asghar Iran-Nejad
Margaret O. Laff
Brian Nash
Theresa Rogers
Terry Turner
Paul Wilson

Abstract

This paper addresses the problem of the effects of prior knowledge, especially those relating to bias, in tests of reading comprehension. Quantitative and qualitative effects of prior knowledge on reading comprehension were demonstrated through an examination of performance on different question types. The availability of the text during question answering was also found to influence performance on certain question types. Peripheral textual items were most sensitive to such influence, central items and scriptal items were least sensitive. Performance on central questions actually improved when readers could not refer back to the text. The biasing effects of prior knowledge were demonstrated both within subjects and between subpopulations (rural and urban). Bias was shown to operate at the level of the individual suggesting that it should be removed at that level, not at the population level. This was achieved by using a content-specific vocabulary test to estimate prior knowledge. This incidentally resulted in a decrease in the bias due to intelligence. A conventional approach to bias removal (collapsing across several text content areas) also removed the bias due to prior knowledge, but at the same time it increased the bias due to intelligence. This latter bias was also found to be increased when readers were able to refer back to the text while answering the questions. Results are interpreted to suggest modifications of current reading comprehension tests and methods of dealing with bias.

Prior Knowledge and Reading Comprehension Test Bias

The basic premise of this paper is that reading comprehension test scores are affected by both an individual's reading comprehension ability and his or her prior knowledge. The main thesis involves a demonstration of the consequences of our inability to distinguish between these two sources of test score variance. A second thesis is a description of a possible solution to the problem.

Prior Knowledge and Reading Comprehension

For many years it has been known that prior knowledge influences what is understood from text (e.g., Bartlett, 1932; Reynolds, Taylor, Steffensen, Shirey, & Anderson 1981). Several studies have suggested that prior knowledge is an integral part of the comprehending process (Bransford & Johnson, 1972; Johnston, 1981). This implies that two individuals equal in reading comprehension ability but differing in prior knowledge would, in all likelihood, exhibit different levels of comprehension of the same text. Such differences are thus likely to show up in assessments of reading comprehension ability, and there is no way of knowing what part of an individual's score is due to reading comprehension ability and what to prior knowledge. Thus attempts to compare several individuals in terms of their reading comprehension ability, are confounded by the differences in their relevant prior knowledge. Findings are then subject to misinterpretation. One student may do very poorly because of a lack of prior knowledge whereas another student, with perfectly adequate prior knowledge, may do poorly because of inadequate reading comprehension

skills. It seems important to distinguish between such sources of failure since each requires quite different assistance.

Test bias is any factor other than that being measured which systematically influences an individual's test score. Prior knowledge constitutes such a factor. The issue is, what to do about the problem. We could try to construct tests which are somehow less dependent on prior knowledge. Alternatively, we could try to obtain an indication of that part of the comprehension score which varies more closely with reading comprehension ability than with prior knowledge, and hence provides a more valid index of raw comprehension ability. The present paper is intended to: (a) show that the former approaches cannot succeed, (b) provide a methodology which may allow us not only to get a less contaminated measure of reading comprehension, but also to distinguish between individuals who fail to comprehend because of prior knowledge mismatches or because of inadequate skill development.

Current Approaches to Test Bias

Existing approaches to reading comprehension test bias all endeavor to devise tests of reading comprehension which are independent of differences in individuals' background knowledge. Three approaches have been used to create such tests: broad topic coverage, passage dependency, and latent trait models.

The first approach is evident in the current tests of reading comprehension which use a number of relatively brief passages each about a different topic. This strategy is based on the idea that diverse text topics ensure that overall, each child gets a similar spread of familiarity

of text. The probable net effect of such a strategy is to ensure that readers with stronger general knowledge will be better prepared for the test of reading comprehension (just as they would be for an I.Q. test or for a vocabulary test).

The second bias reduction method is to eliminate test items which students with extensive prior knowledge could answer before they read the passage. Such questions are called passage (or context) independent (Hanna & Oaster, 1978-79; Tuinman, 1974). If prior knowledge has extensive effects on reading comprehension itself, it is not at all clear that this will solve the problem.

Latent trait theory and related statistical models represent a third potential solution to the problem (e.g., Linn, Levine, Hastings, & Wardrop, 1980). This group of methods is based on statistical theory rather than on a theory of what is causing the bias. Indeed, Tuinman (1979) claims that we have reached the functional limit of mathematical and statistical models, their increased accuracy not being warranted by the accuracy of the actual data. Furthermore, these techniques are based on population-level differences such as skin color. Such population-level approaches seem inadequate for several reasons. In the present context, variability between populations will virtually always be considerably less than the variability between individuals within those populations. In addition, one must make a decision as to which of the many populations to choose as reference groups (e.g., black/white, male/female, urban/rural).

On the larger scale, all of these approaches can be criticized because the basic assumption, that it is possible to construct a reading comprehension test which will produce a score which is immune to the

influence of prior knowledge us erroneous. Since prior knowledge of a topic cannot be equated across readers, we would need to construct a test which was uninfluenced by prior knowledge. Unfortunately, prior knowledge is an integral part of the reading comprehension process (Johnston, 1981; Pearson & Johnson, 1978). Consequently, if test constructors managed to produce a test in which performance was indeed unaffected by prior knowledge, whatever it measured, it would not be measuring reading comprehension.

If it is, as claimed, impossible to construct an umbiased test of reading comprehension, one simply could concede that the test was biased, and obtain a measure of the extent of the bias. The information would be used in the interpretation of the test rather than in its construction. The challenge would be to find a measure of the bias for a given individual. To do this, perhaps we should go to what seems to be the (or at least a major) root of the problem, and look at individual differences in prior knowledge as sources of bias. The question then becomes how to estimate an individual's prior knowledge, and hence the probable test bias for that individual?

Estimating Prior Knowledge

Studies of prior knowledge have generally used "familiar" versus "unfamiliar" texts (e.g., Freebody, 1980) or skin color (e.g., Reynolds, et al., 1981) as estimates of prior knowledge. Two other approaches have also been used. Hagerup-Neilsen (1977) and Raphael (1981) have had subjects rate the familiarity of passages or topics. Unfortunately, aside from the incomparability of different individual's ratings, this procedure

requires metacognitive awareness. Pearson, Hansen, and Gordon (1979) took a more direct approach. These investigators asked eight prior knowledge questions before children read the passages. This seems to be a more powerful approach but the questions tend to over-direct reading. Furthermore, when the questions are highly related to the text, any related improvement could be attributed to greater passage independence of the items. Nonetheless, this more direct approach to the measurement of prior knowledge was used in the present study with modifications which minimize the above problems.

The major problem with any question construction is definitional. Definitions which allow consistent production of other specific item types still elude researchers. It is possible that what is required is a complete theory of the structure of knowledge so that one could generate for any subset of knowledge, appropriate indicators of prior knowledge. However, such a theoretical development is presently unavailable.

A useful set of items should perhaps include some which are very text specific, but these would tend to identify those readers for whom the text contained little, if any, new information. That is, the items would identify those readers for whom reading (that passage) was largely recognition (Tuinman, 1979). Schema theory, however, assumes a more widespread influence of prior knowledge. Consequently, these items alone would be inadequate. Rather, items would need to be symptomatic of relevant underlying schematic knowledge. For example, knowledge of the meanings of certain relatively low frequency words might be diagnostic if the frequency of use was somewhat higher amongst experts in the knowledge domain.

However, items which merely discriminate experts from nonexperts would not be sufficient. A most useful set of items, from a purely functional standpoint, would form a Guttman scale which would differentiate various levels of expertise. This outcome probably would require successively less specific items in order to distinguish the experts from the diletantes, and these from the novices, and so on. In Anderson and Freebody's (1979) terms, we need a spread of items to assess the "depth" rather than the "breadth" of relevant vocabulary. Currently we must take a pragmatic approach to the selection of these items, tempered by such theory as exists. Consequently, in the present study, prior knowledge was measured by testing specific, content-related vocabulary knowledge.

There is, however, a problem with using a vocabulary measure as an estimate of prior knowledge. It would not be difficult to build an argument that vocabulary questions merely estimate general ability (I.Q.) since intelligence tests contain vocabulary subtests. Such tests (and subtests) are highly predictive of performance on tests of reading comprehension. For example, invariably, factor analytic studies of reading comprehension have found a word knowledge factor on which vocabulary tests load highly (e.g., Davis, 1944, 1968; Spearitt, 1972). In studies of readability too, any index of vocabulary difficulty accounts for about 80 percent of the predicted variance (Coleman, 1971).

Anderson and Freebody (1979) have examined the three competing hypotheses which attempt to explain this finding: the instrumentalist, the aptitude, and the background knowledge hypotheses. The instrumentalist position is that knowing words allows text comprehension and not knowing them means that one cannot proceed adequately through the text. The

aptitude hypothesis considers vocabulary knowledge as just another index of I.Q. which is the real factor accounting for comprehension. The background knowledge hypothesis suggests that vocabulary knowledge is a distal index of background conceptual frameworks (schemata) necessary to understand passages about a particular topic.

Although these hypotheses are not mutually exlcusive, the study presented in this paper will test the prior knowledge and general ability hypothesis. That the vocabulary measure estimates prior knowledge and not merely I.Q. will be ensured by a within-subjects design. That is, an individual's I.Q. is relatively stable, thus variability in performance over a two hour period cannot readily be attributed to changes in general ability.

Prior Knowledge and Question Type

The outcome measures from reading comprehension tests generally provide a quantitative measure of "how much the reader has comprehended." There are, however, possible qualitative differences between readers. For example, the total score may be the same for two different readers, but if one succeeded on all literal items and on none of the inferential items, while the other performed equally well on each type, presumably there is a qualitative difference in their comprehension of the text.

Perhaps prior knowledge differentially influences performance on different question types (Pearson, Hansen, & Gordon, 1979). But what constitutes a different type of question? Pearson and Johnson (1978) and Lucas and McConkie (1980) have developed systems which make the same basic distinctions among questions. These distinctions are exemplified in

Pearson and Johnson's system which is really a classification of question-answer relationships. The distinctions relate to the location of the information required to and/or actually used to answer the question. Textually Explicit (TE) items have both the question information and the answer information stated in a single sentence in the text. Textually Implicit (TI) items have the question information and response information stated in different sentences in the text, requiring the reader to combine the separate pieces of information in order to produce or recognize an answer. In order to answer Scriptally Implicit (SI) questions, the reader must combine some information from the text and some from background knowledge (script). Based on the analysis of what is involved in answering the different question types, it seems likely that the SI questions/answers will be more influenced by prior knowledge than will other question types Indeed, Pearson et al. demonstrated this to be so. However, perhaps answering the questions with the text available for reference (as in standardized reading comprehension tests) would produce a different result. For example, since textually implicit questions would then have the reader dependent on memory for neither piece of information, their outcome should become less influenced by prior knowledge.

Of course, prior knowledge may affect other qualitative aspects of the outcome. For example, the reader's performance on more or less central questions may differ depending on his prior knowledge and the extent to which long-term memory is involved in the task. Conceptual dependency theory (Schank, 1975) holds that knowledge is stored with respect to central causal chains of underlying conceptualizations. When readers are

dependent upon their memories for information to answer questions, they are likely to be able to respond more successfully to central items, since central information is more likely to be stored than is peripheral information. However, this may not be the case when long-term memory is only minimally involved in the task, as when the reader can refer to the text while answering questions.

Question classification in tests currently is based around a simple literal versus inferential distinction. Pearson and Johnson's (1978) descriptors represent a more refined version of this approach, yet there is good reason to believe that the "centrality" of the information is also very important. Omanson's (1982) work with the narrative analysis is particularly noteworthy in this regard. It is of considerable theoretical interest to see which set of variables is more important under different task conditions. Pearson and Johnson's descriptors represent the presumed information source, whereas centrality represents more the nature of the information and how it relates to prior knowledge. Once the text has been read and the reader is answering questions from memory, the information source should become less meaningful, since it all must come from the reader's head. However, because of the nature of the storage process, the structural importance of the information is more likely to determine the ability to respond to questions. On the other hand, when text is readily available for referral during question answering (as in standardized tests), it seems likely that location of information (within the text or in the reader's head) should be a much stronger determinant of the reader's responses than the relative centrality of the information. Search strategies may be more critical, and storage should no longer be a problem.

Consequently, the present study used questions based both on Pearson and Johnson's (1978) taxonomy and the centrality notion, to examine possible differential biasing effects of prior knowledge on different types of questions. Similarly, comprehension questions were presented both with and without the text available to refer back to.

It was hypothesized that prior knowledge would account for a significant portion of reading comprehension variance within subjects, thus representing an important biasing factor. It was anticipated that the biasing effects would not be accounted for on the basis of the passage dependency of the questions and neither would the problem be removed by increasing the spread of text topics. Instead, increasing the spread of text topics was expected to increase the correlation between total reading comprehension score and I.Q. However, it was predicted that bias would be removable by estimating prior knowledge with a content-specific vocabulary test and producing residual comprehension scores.

The effects of prior knowledge were also hypothesized to differ across question types depending on whether or not the text was available to refer back to while answering the questions.

METHODOLOGY

The Materials and Tasks

Reading Comprehension

Reading comprehension was assessed by having the students read and answer 18 questions about each of three 650-750 word texts. The content areas of the texts were:

(1) The specialization of corn in the U.S.

(2) The financial problems of the Chicago Regional Transit Authority (RTA).

(3) The battle of Antietam Creek.

The first two topics were chosen for their likely bias toward rural and city children, and the third for its presumed lack of bias (since the Civil War is part of both groups' curricula). The Fry readability scores of these texts were seventh grade (Civil War) and eighth grade (corn and RTA). The texts were basically taken from a textbook (Civil War), an agriculture handbook (corn), and two newspaper articles (RTA).

The 18 questions were constructed for each text with 6 of each type of question in Pearson and Johnson's (1978) taxonomy: textually explicit, textually implicit, and scriptally implicit. In addition, half of the items for each question type tested information which was central to an understanding of the text and half tested peripheral information. These divisions were accomplished by having ten adult subjects rate on a 1-4 scale the centrality of a list of propositions derived from the passages. Propositions were considered to be central if the mean rating was three or higher, and peripheral if two or lower. This criterion generally meant that there was at least 80% agreement among the adults in whether the item was given one of the top two or bottom two ratings. The selected propositions were then turned into multiple-choice questions by generating alternatives such that two of the distractors maintained some of the surface characteristics of the text. Each set of questions thus contained three of each of the six question/answer types generated by the Pearson and Johnson classification system and high versus low centrality.

A problem occurred which related to the nature of the Pearson-Johnson taxonomy. Unless textually explicit or implicit questions and answers are verbatim from the text, they involve varying amounts of scriptal knowledge. That is, as soon as a synonym is substituted, scriptal knowledge becomes mildly implicated in the relationship. In the present study, synonym substitution or paraphrase was allowable within textual items. Scriptal items required an extra piece of information which was not mentioned in the text.

Prior Knowledge--Vocabulary Tests

The extent of an individual's prior knowledge relevant to each of the content areas used in the reading comprehension test passages was assessed by means of content-specific vocabulary questions. Each of the three content areas was addressed with 11 multiple-choice questions, each presenting a word and four possible definitions, or a definition and four possible words. The 33 items were placed in a single test format, with the content areas alternating so that every third question addressed the same content area. The resulting general vocabulary test contained three content-related subtests. The vocabulary which was assessed by the questions was selected so that some items were very specific to the content area, whereas other items were somewhat less specific. This was done in an effort to distinguish varying degrees of "expertness." In the present study, this specificity was done at an intuitive level.

When the vocabulary test was administered, the students were simply told that the test was a vocabulary test and that they were to work through

it at their own pace. They were also told how to answer the questions (circle the correct alternative) and to be sure to answer all questions.

Intelligence Test

As a measure of intelligence, the students were given the IPAT Culture Fair Intelligence Test scale 2 (Institute for Personality and Ability Testing), a nonverbal reasoning test involving four subtests and taking about 20 minutes to administer.

Subjects

A total of 207 eighth-grade students from two quite distinct subpopulations participated in the study: Three small rural schools in southern Illinois (N = 101), and two parochial schools in Chicago (N = 106). The mean I.Q. on the IPAT culture-fair was 103 (SD = 14.5) with subpopulation means of 101.01 (SD = 13.94) for rural students, and 104.83 (SD = 14.89) for urban students.

Procedure

In order to ensure that ability was equally spread across the groups, scores on standardized reading comprehension tests were obtained several days prior to the study and were used to rank order students before assigning them to groups, thus producing stratified random samples.

There were four between-subject experimental conditions. Three of these were based upon the extent to which subjects were dependent upon long-term memory to answer the questions. Group One (N = 45) was least dependent on long-term memory since it had the text available to refer to while answering the questions. Group Two (N = 47) was not allowed to look

back at the text while answering the questions, but proceeded to answer the questions as soon as the passage was read. The third group (N = 49) was not only unable to refer back to the text while answering the questions, but had a five-minute task interposed between reading a text and answering the questions. The tasks used were subtests of the IPAT non-verbal which the other groups took in one sitting.

The fourth group (N = 50) was a control group. These students were required to answer the questions without the benefit of having read the text. Such a group was necessary in order to demonstrate that the effects of prior knowledge were not simply on question answering, but on reading. In each school, Group Three was tested separately from Groups One, Two, and Four since only they required systematic interruption of their reading and question answering.

Each student was given an envelope containing the necessary materials. All took the vocabulary test first. Groups One, Two, and Four then took the IPAT non-verbal I.Q. test followed by their comprehension tests. The third group received their texts in a different manner. They were given the text, then a section from the IPAT, followed by the questions. This pattern was then repeated for each of the other two text topics.

Results and Discussion

All major analyses involved split plot hierarchical multiple regressions (Cohen & Cohen, 1975). Since the within-subjects measure of prior knowledge was not independent of the between-subjects measures, the individual's mean score on the dependent variable was entered as the first independent variable in the within-subjects analysis. This procedure has

the effect of removing all between-subject variance and leaving only within-subject variance (Erlebacher, 1977).

All students read the passages in the same order, and the passages were clearly not of equal difficulty. These effects were removed by entering "passage" (as two orthogonal contrasts) second in the within-subjects analysis. Since there was no reason to hypothesize equal (or unequal) difficulty of the passages, these usually significant contrasts were not interpreted.

If a subject skipped a page of questions, then those data were labeled missing. However, an omission of one or two questions in sequence resulted in the items being marked incorrect. Only subjects with complete data were used in the analyses.

The Experimental Tasks

Reading Comprehension

A problem arose with the comprehension task. While the readability of the texts was rated at the seventh and eighth grade difficulty by the Fry formula, the students' comprehension scores indicated that the task was very difficult. Of course, rather than the texts, the problem may have been more in the questions. Indeed, for about five of the questions on each text, the students' mean response was at or below chance level. The effect of this "flooring" was to produce a restriction of range. Nonetheless, rather than tamper with the data by discarding these items, it was decided to analyze the intact data. The findings must be interpreted in the light of this range restriction, and the question of possible underestimation of effect sizes must be considered.

The Prior Knowledge-Vocabulary Tests

This set of tests functioned well, having a full range of scores (1-11) on two tests, a range of 2-11 on the third, and means of 8.1 (corn), 6.4 (RTA), and 6.7 (Civil War). Standard deviations were 1.9, 2.0, and 2.2, respectively.

Prior Knowledge and Reading Comprehension

A major focus of this study was an investigation of the effects of prior knowledge upon reading comprehension. Three different observations were taken on each variable for each subject, one for each knowledge domain. This means that if prior knowledge differences influence reading comprehension for a given individual, then it is difficult to argue that the effects were due to some other factor such as verbal I.Q., which would be constant for that individual.

Because the within-subjects design really does allow the "all else being equal" assumption in interpretation, one should not expect as much variability within subjects as exists between them. However, one can expect effects which are less contaminated by extraneous variables. Furthermore, the number of observations involved in the within-subject side of this study is three times that for the between-subjects side. Since the analysis is consequently less likely to "overfit" the data (that is, repeated samples are likely to yield very similar findings), the variables tend to explain less dramatic but more reliable proportions of variance.

The Findings

Reading Comprehension and Test Bias

The first major finding of the study was that prior knowledge accounted for 3.5% of the within subject variance, $F(1,282) = 11.72$, $p < .001$, Table 1. This result indicates that prior knowledge influences the

---

Insert Table 1 about here.

---

comprehension of texts independent of the effects of intelligence and other between-subject confounding variables. The evidence cannot be argued on the grounds of contrived materials or other validity grounds since it has been replicated with a selection of very ordinary texts, and using multiple-choice questions. The potential of prior knowledge as a biasing factor is evident.

The study also offers insight into the practical implications of this biasing effect for the assessment of reading comprehension. Between-subject variability shed most light on this issue. While the proportions of variance explained are inflated by reduced degrees of freedom (though still substantial) and a greater possibility of correlated nuisance variables, between-subject variability reflects the assessment situation more accurately. The proportions of between subject variance accounted for by prior knowledge, with general ability held constant, are shown in Table 2. The effect is consistent across texts.

Insert Table 2 about here.

The effect was not simply due to readers' ability to answer the questions regardless of having read the text. This possibility was investigated through a regression analysis of the scores of students who answered the questions without having read the text. The proportions of between subject variance explained by prior knowledge for each passage were 2% (corn), 1% (city), and 4% (Civil War), none of which was significant ($\underline{N}$ = 50). Consequently, attempts to remove bias by simply discarding the less text-dependent items seem unlikely to succeed.

To see whether the texts were in fact biased towards one or another subpopulation, reading comprehension scores were regressed on prior knowledge and the subpopulation of which the reader was a member (in both orders). Table 3 shows that the texts used were each biased towards either the rural or the urban children (population entered first). The "corn" passage was biased toward rural students, and the "city" passage was biased towards urban children. These biases had been predicted a priori, but the "Civil War" passage (presumed to be neutral) was also biased towards rural students. Possibly the country children's curriculum covered more (or more relevant) Civil War material.

Insert Table 3 about here.

While this demonstrates that population level bias exists, bias is not a population level phenomenon but an individual one. Two findings support this claim. First, there was a trend towards a sex bias in the "Civil War" passage. Boys tended to know more about war things and to read about them with greater comprehension. While not statistically significant, $\underline{F}(1,139)$ = 3.71, $\underline{F}$ required for significance at .05 level = 3.91; this trend

illustrates the fact that when bias is defined at the population level, there are potentially as many biases as we can describe subpopulations. Second, when prior knowledge is entered into the regression before subpopulation, the latter has virtually no remaining predictive power. Thus, removal of the population level bias can be accomplished by removing the individual level bias, but the reverse generally is not true.

There are two ways to examine systematic effects, and each is represented by one of the above definitions. An empirical demonstration of group differences represents the current definition. However, there are as many such potential biases as there are conceivable subpopulations. Most group biases normally go unnoticed simply because we lack the population descriptors and motivation to test for them. It is because of this that we cannot simply try to statistically identify biased items and then eliminate them from the test post hoc. How many subpopulation descriptors should we use? Just the politically expedient ones?

The second way to examine systematic effects is through theory. If we have a theory of the source of biases, we can look at bias at the individual level. The proposed definition recognizes that a test can be biased against an individual within a population. Identification of such bias need no longer be dependent on differences between arbitrarily selected subpopulations. Theory offers us a solution to the problem of test bias. The solution involves adopting an approach not unlike that commonly taken over the I.Q./reading comprehension relationship. That is, initially it has been accepted that reasoning <u>is</u> an integral part of reading (Johnston, 1983; Thorndike, 1917; Tuinman, 1979); thus nobody

tries to construct reasoning-free reading comprehension tests. Instead, they are satisfied examining reading comprehension in the context of a measure of reasoning ability such as a WISC score. Perhaps the same should be done with a measure of prior knowledge. This study shows that having measured relevant prior knowledge its effects can be removed statistically from tests of reading comprehension when required. Removing the effects of prior knowledge provides us with a residual reading comprehension score which is free from bias.

There are several criticisms which might be leveled at this approach. It might be protested that the prior knowledge bias can be eliminated more easily by using a variety of text topics to produce an aggregate score, as is done in current tests. Table 4 shows that indeed this is the case. However, the figures also indicate that there is an unfortunate side effect of such a procedure. The proportion of variance related to I.Q. becomes much greater. That is, an I.Q. bias has been introduced. On the other hand, the population difference also disappears when the bias is removed statistically from each passage score before aggregating the "debiased" scores. But there is also a beneficial side effect. The extent to which I.Q. explains performance is also reduced considerably, from 14.6% of the variance to 4.1%. This reduction is significant at the .001 level using a dependent sample t test for differences between variances, $t(138) = 20.43$. Furthermore, the table illustrates what may be measured by reading comprehension tests. Removing the influence of prior knowledge leaves a variance of 6.15 instead of 35.4, 17% of the original variance. In other words, 17% of the variance in the measure is due to factors which are independent of prior knowledge.

Insert Table 4 about here.

Critics may well question the reliability of residual scores. Substantial norming populations and well designed tests may reduce this problem somewhat. However, it must be born in mind that current methods are no better. Any greater reliability of our scores on conventional test scores is not due to their reliably measuring reading comprehension, because a good part of the raw score is a result of differences in intelligence and other factors. Thus, the greater raw-score reliabiity is at the expense of validity.

Critics might wonder about the context in which a residual score might be useful. In order to address this issue it is important to make a distinction between the use of the prior knowledge measure at the individual level and at the group level. The residualized score is most useful at the group level where one is interested in knowing how able one or more groups of readers are at comprehending from text given their levels of relevant prior knowledge. Interestingly, when the debiased scores for individual passages are summed into a total score, the net score is not only free from prior knowledge bias, but also relatively free from general reasoning bias (Table 4). Both effects are because we have removed the cause (rather than just the symptom) of the biases from the test. With the cause gone, the symptoms go too.

Alternatively, the effects need not be removed. Instead, performance on the comprehension test might simply be considered in light of a measure of the reader's prior knowledge. In this case, with appropriate norms, a

reader's performance might be considered separately on familiar and unfamiliar material. Since there are different strategies involved in reading familiar and unfamiliar texts (also depending on the reader's goal), such evaluation may yet provide valuable diagnostic information. At the individual level, the residual score is still meaningful in that it describes the individual's reading comprehension performance relative to that which would be expected given his or her level of prior knowledge. However, when working with individuals, it would be best to have all three scores available for interpretation: the raw reading comprehension score, the prior knowledge score, and the residualized reading comprehension score.

Other types of reading problems ultimately may also be detected using this approach. One such diagnosable reading difficulty may be that described by Spiro (1980) as a "schema selection" problem. This is the problem caused by failure to use relevant prior knowledge when it would be appropriate to do so and the reader has it available. Of course, problems caused by "schema unavailability" would also be readily detected, that is, failures caused simply by the reader not having the appropriate relevant knowledge base before reading. While these proposals remain, for the moment, untested, the promise is great, and they are an important area to be developed in future research. The first step towards this must be the refinement of the measure of prior knowledge.

Reading Vocabulary and Reasoning

Anderson and Freebody (1979) have described three hypotheses to explain why vocabulary tests account for so much of the variance in reading

comprehension tests. The first of these hypotheses is the "general ability hypothesis." This hypothesis proposes that the relationship is simply that vocabulary tests estimate general ability and brighter students will be better readers. This study provides evidence against this hypothesis. First, the within-subject analysis involving the prior knowledge vocabulary test shows the effect of prior knowledge on comprehension (Table 1). Since it does not seem reasonable to assume that an individual's general ability varies from moment to moment, these effects do not support the general ability hypothesis.

Second, in the between-subject analyses (Table 2), the variance associated with reasoning ability (as measured by the IPAT) was covaried out before the prior knowledge vocabulary scores entered the regression equation. The prior knowledge test still accounted for a substantial portion of reading comprehension variance. Thus, at least some of the relationship between vocabulary and reading comprehension is not simply because both relate to general ability.

While these findings argue against the general ability hypothesis, they support the "prior knowledge hypothesis" which asserts that the connection between vocabulary and reading comprehension tests is prior knowledge. That is, knowing the words in the vocabulary test is indicative of underlying schemata. At least this is so in the single text situation. Standardized tests, however, use more than one text.

Contemporay reading comprehension tests contain a number of texts each on a different topic. Vocabulary tests also contain items from a broad range of domains. Combining the content-specific vocabulary tests into a single nonspecific vocabulary test would reflect this situation and at the

same time produce a longer more reliable test. If general verbal ability is the source of the relationship between current vocabulary tests and reading comprehension tests, a more general vocabulary test should correlate more highly with comprehension of a given passage. To test this hypothesis, three vocabulary scores were constructed for each passage as follows:

(1) the sum of the 11 content-specific items (specific vocabulary)

(2) the sum of the remaining 22 items (general vocabulary[2])

(3) the sum of all 33 items (general vocabulary[3]).

Insert Table 5 about here.

The mean correlations between these three scores, I.Q., and reading comprehension (Table 5) suggest that the more vocabulary tests are aggregated across content, the more they correlate with I.Q. and the less with reading comprehension, though the trend is not statistically significant.

It could be argued that this relationship with I.Q. is simply because of increased reliability as a result of more test items being aggregated. To counter this argument, two similar general vocabulary tests were constructed, each containing a random sample of items with the restriction of equal numbers from each content area instead of all items from each specific test. This provided three tests of differing generality but with equal numbers of items. Table 5 shows in parentheses the correlations between these tests, reading comprehension, and I.Q. The figures suggest that the increased correlation with I.Q. is due more to increased diversity

of content than to increased reliability. Vocabulary tests with equal numbers of items but increasing generality were still increasingly correlated with I.Q.

Further support was gained for this hypothesis, by entering the general vocabulary[2] scores into the regression before the specific vocabulary. If the prior knowledge hypothesis is correct, the specific vocabulary test should still account for a significant proportion of variance in reading comprehension, even after the statistical removal of the effects of the general vocabulary[2] test. This was indeed the case. The 22 item general test accounts for an average of 3.9% of the reading comprehension variance whereas the specific test accounts for an average of 9% of the variance (Table 6). This finding is in spite of the fact that the general test has twice as many items, covers a broader span of knowledge, and enters the regression first.

Insert Table 6 about here.

Table 4 presents a different perspective on the problem. When the effects of prior knowledge are removed from each passage, and the individual's total residual score is computed, I.Q. accounts for a very much smaller portion of the variance than it does when the raw (biased) scores are aggregated. It is still significant, as one would expect (Johnston, 1983; Thorndike, 1917; Tuinman, 1979), but explains a smaller proportion of the variance.

From these arguments, it can be seen that while the prior knowledge hypothesis is supported for specific vocabulary and comprehension of specific texts, the standardized tests provide a situation best described

by the "general ability" hypothesis. Aggregating performance on vocabulary or reading comprehension tests across content areas tends to increase the correlations between those tests and tests of I.Q. because both are biased towards greater general knowledge.

A further source of relationship between standardized reading comprehension tests and I.Q. was also explored. It was suggested in the first section of this paper that part of the correlation between I.Q. and reading comprehension in standardized tests may stem from the fact that the text is fully available for the reader to refer to for answers. Such tests require search and match strategies, this hypothesis was testable.

Indeed, the hypothesis did gain some support from the correlations between I.Q., comprehension; and prior knowledge when the task depends increasingly on long-term memory. When the text is available the correlation between I.Q. and comprehension is higher ($\underline{r}$ = .31) than when the text is not available but questions are immediate ($\underline{r}$ = .27) which is, in turn, higher than the correlations when the text is unavailable and the questions are delayed ($\underline{r}$ = .19). The reverse trend is evident for the correlations between comprehension and prior knowledge. When the text is available the correlation between prior knowledge and reading comprehension is lower ($\underline{r}$ = .23) than when the text is not available but questions are immediate ($\underline{r}$ = .24) which is lower than when the questions are delayed ($\underline{r}$ = .33). While these correlations are not significantly different from one another, they consistently proceed in opposite directions as predicted. The probability of these two trends occurring by chance is .063. Because the two trends procede in opposite directions, it is difficult to argue

that the reduced correlation with I.Q. might be due to reduced variance or some other alternative. Thus, the data suggest that standardized reading comprehension tests are biased towards readers with greater general ability.

Question Type and Long Term Memory Demands

The effects of prior knowledge on reading comprehension when the test tasks made readers more or less dependent on information storage and retrieval were examined using three groups of subjects.
Group One subjects had minimal dependence on memory since they had full access to the text while answering the questions. Group Two was denied such access to the text but answered the questions as soon as they had read the text. The third group was denied text access during question answering and had an interfering task between text and questions.

The contrast between groun one and the other two groups was significant, $F(1,282) = 7.67$, $p < .01$. The means for the three groups were 7.4, 6.3, and 6.2, respectively (standard deviations 3.0, 3.0, 2.8). The contrast between the latter two groups was not significant, possibly because of floor effects, and possibly because the approximately five minute filled delay was not long enough to induce further changes in performance. However, the major interest in this variable was in its relative effects on different question types.

For the analysis of the effects of different question types, each subject's comprehension score was broken down, within each topic, into six subscores, representing the three question types by two levels of importance. Importance was dichotomously coded and question type was

entered into the regression as two orthogonal contrasts: Q1 representing the contrast between textual items (the mean of the textually explicit and textually implicit items) and scriptally implicit items; Q2 representing the contrast between textually implicit and textually explicit items. The results of this analysis are shown in Table 7.

Insert Table 7 about here.

Each subscore contained only three multiple choice items. Consequently, the subscores have a high error component and a very small variance which was restricted further by the generally low performance. These constrictions are reflected in the proportions of variance explained. The proportions should be given less credence than the F values.

Both question type contrasts were significant, reflecting the fact that textually explicit questions (mean = 45%) were easier than the textually implicit questions (mean = 37%) which were easier than the scriptally implicit questions (mean = 29%). As a main effect, centrality of the piece of information being assessed was not a significant predictor of performance. However, both centrality and question type are involved in significant interactions with other variables.

Two series of interactions were significant. The first of these involved prior knowledge, centrality, and the availability of the text while answering questions. When the text is available to refer to, answering peripheral questions is easy (Figure 1). When the text is not available to refer to, the same task becomes very difficult. It seems that peripheral information is easily obtained from searches of the text but less readily stored.

Insert Figure 1 about here.

On the other hand, central questions posed an easier task when the text was not available for reference than when it was available. Schema theory would predict that there should be minimal deterioration in performance on central questions when memory must be relied upon more, since the reader presumably constructs a central chain in the process of comprehending. The fact that performance actually gets better may be because of a preoccupation, on the part of the reader, with the textual features. That is, when the text is available, a reader may use search strategies rather than comprehension strategies. Text based distractors may then prove to be more attractive, since the search would also turn up bits of information found in the distractors. This interpretation is supported by the results of a study by Nicholson, Pearson, and Dykstra (1979) who found that when readers were allowed access to the text (which contained embedded errors) while answering questions, they were less accurate in their answers than if they did not have access to the text. In the present study it is also noticeable that the improvement on central questions is greater for students with greater prior knowledge (Figure 1). This might also be expected if readers were indeed able to more successfully store the central chain of information than the peripheral details.

In addition, Figure 1 shows an interaction between prior knowledge and the centrality of the questions. It indicates that when readers are reading more familiar material they are more able to answer questions about

the text, and this advantage is greatest for more peripheral questions. This can be interpreted in terms of a model which suggests that when readers have greater prior knowledge, they have more highly developed schematic structures, with more accessible "slots" for storing related information. Thus, while the performance on central items does improve, the improvement is more marked on the peripheral questions. In the same way, when readers have little prior knowledge, the biggest decrement in performance when memory is called upon is on the peripheral items. Readers generally answer central questions better when the text is unavailable than when it is available, and this trend is more pronounced when readers have greater topic-relevant knowledge.

The second series of interactions includes those involving centrality, text availability, and question type (the contrast between scriptally cimplicit items and the mean of the two textual items). The contrast between question type and text availability (Figure 2) indicates that while textual questions are easier than scriptal questions, when the reader does not have access to the text the drop in performance on textual questions is extreme. The fact that this falloff in performance is not as severe for the scriptal items is probably at least partly due to an obvious floor effect.

Insert Figure 2 about here.

While central questions are more difficult than peripheral ones, if they are scriptal as well as central, they are even more difficult. Again, the scriptal questions show an improvement when readers do not have access to the text, possibly reflecting their reluctance, when the text is

present, to use their prior knowledge. This may also reflect increased attractiveness of the text-based distractors through the readers' greater belief in the text than in their own understanding. Again, this supports the findings of the Nicholson, Pearson, and Dykstra (1979) study noted above.

The most interesting aspect of this interaction involves the difference between central and peripheral text-based questions across tasks differing in long-term memory demands. The readers' performance on more central textual questions is relatively unaffected when the text is unavailable to refer back to, whereas their performance on peripheral textual questions shows a precipitous drop. This is exactly as could be predicted from the type of general model proposed by Schank (1975) and Kintsch and van Dijk (1978) and that proposed by Omanson (1982) for narrative text.

Summary and Conclusions

This study provides evidence that prior knowledge influences the comprehension of texts and that the effect is not because of contrived materials, or other validity problems. Neither is it simply because of improved ability to guess the questions without first reading the text. This means that prior knowledge can be responsible for biasing the information gained from reading comprehension tests. The study also raises the question of what standardized reading comprehension tests measure. The answer, as indicated by this study, is that they provide a fairly good proxy for I.Q., just as do standardized vocabulary tests. A high score on such a reading comprehension test indicates that the student will probably have little trouble in school, particularly in reading, and that he or she

seems to have the adequate, and appropriate fund of general knowledge expected of a middle-class American student.

What, then, does a low score indicate? This is a much more difficult question. It might indicate that the child cannot read adequately. It might also indicate that his or her store of prior knowledge in the areas tapped by the test is not adequate for the task. Or the student might have, as Thorndike (1917) claims, generally meager processing skills. The question of what to do about the student's problem then arises. Without being certain of the cause, it is very difficult to decide on a course of remedial action.

The study suggests some potential antidotes to the problem. First, if comprehension is defined as the forming of a coherent cognitive model of the text meaning, then interest is most likely to be on the reader storing the central aspects of the text. It seems that the best way to evaluate this is to ask central questions, and possibly to prevent the reader from referring to the text while answering the questions. Note that asking central questions implies that the text should be long enough and structured enough to have a central thread.

There may also be arguments for other question types which might supply diagnostic information. For example, if the definition of reading comprehension includes the use of prior knowledge in constructing the model of meaning, or the integration of the model of meaning with prior knowledge, then it might be useful to ask scriptally implicit questions also, since they require the reader to use prior knowledge. However, in asking such questions it must be recognized that they describe something

different about the reader's comprehension from that which textual questions describe.

By looking at performance on textual and scriptal items in the context of a prior knowledge score, it might be possible to diagnose schema selection problems. The prior knowledge measure by itself enables diagnosis of schema availability problems, i.e., lack of prior knowledge preventing adequate processing of the text. However, the diagnostic aspects of question type have only been scratched by this study. Much more work is needed to develop these question types into systematic and meaningful diagnostic instruments.

The present study demonstrates that prior knowledge is a powerful source of test bias. It has been shown (Johnston, 1981) that the extent of an individual's prior knowledge influences the basic cognitive processes which are involved in reading comprehension. It has also been argued amply and demonstrated elsewhere that prior knowledge influences the inferences which people make as they comprehend text (e.g., Anderson, Reynolds, Schallert, & Goetz, 1976; Spiro, 1975).

The important things to note are that (a) these systematic influences are described at the individual level, not at the population level, and (b) prior knowledge is an integral part of reading comprehension. The consequence of these two facts is that since no two individuals will have identical prior knowledge, the construction of tests which are free of bias at the individual level is impossible. Furthermore, it can be argued that it would be undesirable in any case since a reading comprehension test uninfluenced by prior knowledge would certainly not be measuring reading comprehension as it is understood theoretically.

At the level of standardized achievement tests, a major advantage of an approach which involves measuring prior knowledge has been demonstrated in the present paper. Bias can be effectively removed from tests by partialing out the effects of prior knowledge. The valuable aspect of the bias removal is that it is not a widely recognized bias. Indeed, it shows that there are probably many biases, since bias arises at the individual level, not at the group level. The proposed approach allows us to avoid the dilemma of which group biases to attempt to remove.

The proposed method of bias removal has a further advantage. Since reading comprehension involves not only being able to locate specific information on a page, but forming a coherent integrated representation of the information, more substantial text segments are called for. The introduction of prior knowledge measures would allow this luxury since it would no longer be necessary to increase the number and variety of texts to reduce bias. Few would deny the greater validity of comprehension estimates based on more substantial segments of text. Apart from the greater flexibility which they allow in terms of question generation, longer texts allow more structure to be built into them and they have greater ecological validity.

Furthermore, since forming a coherent representation is almost unnecessary when the text is available to return to while answering the questions, perhaps at least some parts of tests should not allow such access. This may provide a better assessment of understanding of the central aspects of a text since variability is associated more with central than peripheral questions in the no access situation. Knowing that

the two types of task require different skills, particularly given differing prior knowledge, it may be possible to form a better judgment as to the cause of a child's reading problem. Note, too, that these advantages hold whether the test is for diagnostic or for survey purposes. While the approach does provide information which is diagnostic, when used for bias removal, this information can increase the construct validity of the test score, since it provides an estimate of reading ability which is less contaminated than current test scores by differences in prior knowledge and general ability.

Such advantages are not restricted to the standardized test arena. The classroom teacher, and other informal assessors (reading specialists, etc.), can also accomplish the same task with a few well chosen questions. Indeed, most teachers already ask relevant prior knowledge questions as a prelude to reading, largely as a "schema activation" procedure, to help the students bring their knowledge to bear on the text. These same questions can serve the dual function of alerting the teacher to the nature and extent of the children's relevant knowledge, thus providing an insight into the nature of the task demands upon the students.

It is important that educators begin to look at comprehension skill in the context of the students' relevant prior knowledge, a suggestion made feasible by the finding that a brief content-relevant vocabulary subtest can provide a reasonably good indicator of prior knowledge. This use may be most obvious in assessments of reading in the content area. Unless the prior knowledge measure is available, little can be said about a student's ability to read content area text. Failure may be due to inadequate prior knowledge, inadequate strategies, or both. Sternberg (1981) claims that in

mental testing the diagnostic goal is to be able to decide whether various processing components are unavailable, inaccessible, or inefficiently executed, and whether the components and strategies operate on an inadequate mental representation. He suggests that perhaps "cognitive contents" tests are needed as well as cognitive components tests so that both knowledge and processing deficiencies can be assessed. Clearly there is room for improvement on the test questions developed in this study, but by the systematic examination of various domains, the ability to construct such tests should improve considerably.

The data on question types also suggested the possibility of a reliability-validity tradefoff in current assessment procedures. When the text is available for reference while answering questions, the item type which distinguishes best between high and low knowledge readers is the peripheral item. Consequently, if items are selected on the basis of the discrimination index, we will end up with tests which tend to be composed of relatively trivial items just as Tuinman (1979) suggests. Indeed, Johnston and Afflerbach (Note 1) have provided evidence that such is the case. Is this what we wish to measure? Is it really what we consider to be comprehension? We must begin to look carefully at our priorities on these issues. A deeper understanding of exactly what we are getting from our current measures and of the alternatives should help in this matter.

In conclusion, this paper was motivated by disenchantment with the assessment approach of controlling "nuisance" variables, particularly prior knowledge, by randomization. The approach cannot work. In particular, bias cannot be eliminated by collapsing across various content domains and

throwing out items which violate an expected distribution. A more productive approach is to measure such "nuisance" variables and take them into account, as the valuable information which they are, for our assessment interpretation.

Reference Note

1. Johnston, P., & Afflerbach, P. The centrality of reading comprehension test questions and their validity. Paper presented at the annual meeting of the New York State Reading Association, Kiamesha Lake, N.Y., November 1982.

References

Anderson, R. C., & Freebody, P. Vocabulary knowledge (Tech. Rep. No. 136). Urbana: University of Illinois, Center for the Study of Reading, August 1979. (ERIC Document Reproduction Service No. ED 177 480)

Anderson, R. C., Reynolds, R. E., Schallert, D. L., & Goetz, E. T. Frameworks for comprehension discourse (Tech. Rep. No. 12). Urbana: University of Illinois, Center for the Study of Reading, July 1976. (ERIC Document Reproduction Service No. ED 134 935)

Bartlett, F. C. Remembering. Cambridge, Mass.: University Press, 1932.

Bransford, J. D., & Johnson, M. K. Contextual prerequisites for understanding: Some investigations of comprehension and recall. Journal of Verbal Learning and Verbal Behavior, 1972, 11, 717-726.

Cohen, J., & Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, N.J.: Erlbaum, 1975.

Coleman, E. B. Developing a technology of written instruction: Some determinants of the complexity of prose. In E. Rothkopf & P. Johnson (Eds.), Verbal learning research and the technology of written instruction. New York: Columbia University Teachers College Press, 1971.

Davis, F. B. Fundamental factors of comprehension in reading. Psychometrica, 1944, 9, 185-197.

Davis, F. B. Research in comprehension in reading. Reading Research Quarterly, 1968, 3, 499-545.

Erlebacher, A. Design and analysis of experiments contrasting the within- and between-subjects manipulation of the independent variable. Psychological Bulletin, 1977, 84, 212-219.

Freebody, P. Effects of vocabulary difficulty and text characteristics on children's reading comprehension. Unpublished doctoral dissertation, University of Illinois, September 1980.

Hagerup-Neilsen, A. R. The role of macrostructures and linguistic connectives in comprehending familiar and unfamiliar written discourse. Unpublished doctoral dissertation, University of Minnesota, 1977.

Hanna, G. S., & Oaster, T. R. Toward a unified theory of context dependence. Reading Research Quarterly, 1978-79, 14, 226-243.

Johnston, P. Reading comprehension assessment: A cognitive basis. International Reading Association, Newark, Del., 1983.

Johnston, P. Prior knowledge and reading comprehension test bias. Unpublished doctoral dissertation, University of Illinois, August 1981.

Kintsch, W., & van Dijk, T. A. Toward a model of text comprehension and production. Psychological Review, 1978, 85, 363-394.

Linn, R. L., Levine, M. W., Hastings, C. N., & Wardrop, J. L. An investigation of item bias in a test of reading comprehension (Tech. Rep. No. 163). Urbana: University of Illinois, Center for the Study of Reading, March 1980. (ERIC Document Reproduction Service No. ED 184 091)

Lucas, P. A., & McConkie, G. W. The definition of test items: A descriptive approach. American Educational Research Journal, 1980, 17, 133-140.

Nicholson, T., Pearson, P. D., & Dykstra, R. Effects of embedded anomalies and oral reading errors on children's understanding of stories. Journal of Reading Behavior, 1979, 11, 339-354.

Omanson, R. C. The relation between centrality and story category variation. Journal of Verbal Learning and Verbal Behavior, 1982, 21, 326-337.

Pearson, P. D., Hansen, J., & Gordon, C. The effects of background knowledge on young children's comprehension of explicit and implicit information (Tech. Rep. No. 116). Urbana: University of Illinois, Center for the Study of Reading, March 1979. (ERIC Document Reproduction Service No. ED 169 521)

Pearson, P. D., & Johnson, D. D. Teaching reading comprehension. New York: Holt, Rinehart & Winston, 1978.

Raphael, T. E. The effect of metacognitive awareness training on students' question answering behavior. Unpublished doctoral dissertation, University of Illinois, 1981.

Reynolds, R. E., Taylor, M. A., Steffensen, M. S., Shirey, L. L., & Anderson, R. C. Cultural schemata and reading comprehension (Tech. Rep. No. 116). Urbana: University of Illinois, Center for the Study of Reading, March 1979. (ERIC Document Reproduction Service No. ED 169 521)

Raphael, T. E. The effect of metacognitive awareness training on students' question answering behavior. Unpublished doctoral dissertation, University of Illinois, 1981.

Reynolds, R. E., Taylor, M. A., Steffensen, M. S., Shirey, L. L., & Anderson, R. C. Cultural schemata and reading comprehension (Tech. Rep. No. 201). Urbana: University of Illinois, Center for the Study of Reading, April 1981.

Schank, R. C. The structure of episodes in memory. In D. g. Bobrow & A. Collins (Eds.), Representation and understanding: Studies in cognitive science. New York: Academic Press, 1975.

Spearitt, D. Identification of subskills of reading comprehension by maximum likelihood factor analysis. Reading Research Quarterly, 1972, 8, 92-111.

Spiro, R. J. Inferential reconstruction in memory for connected discourse (Tech. Rep. No. 2). Urbana: University of Illinois, Center for the Study of Reading, October 1975. (ERIC Document Reproduction Service No. ED 136 187)

Spiro, R. J. Schema theory and reading comprehension: New directions (Tech. Rep. No. 191). Urbana: University of Illinois, Center for the Study of Reading, December 1980.

Sternberg, R. J. Testing and cognitive psychology. American Psychologist, 1981, 36, 1181-1189.

Thorndike, E. L. Reading as reasoning: A study of mistakes in paragraph reading. Journal of Educational Psychology, 1917, 8, 323-332.

Tuinman, J. J. Determining the passage-dependency of comprehension questions in 5 major tests. Reading Research Quarterly, 1974, 2, 207-223.

Tuinman, J. J. Reading is recognition--When reading is not reasoning. In J. C. Harste & R. R. Carey (Eds.), New perspectives on comprehension (Monograph in Language and Reading Studies No. 3). Bloomington: Indiana University, 1979. Pp. 38-48.

Table 1

Partitioning of Reading Comprehension Variance

and Tests of Significance

Variable	*F*	Increment in Percentage of Variance Explained
	Between	
IQ	19.52****	11.91
Text Availability	7.67**	4.68
Question Delay	<1	.20
IQ x Text Availability	<1	.05
IQ x Question Delay	<1	.16
	Within	
Passage Contrast 1	20.62****	6.09
Passage Contrast 2	23.84****	7.04
Prior Knowledge	11.72****	3.46
IQ x Prior Knowledge	<1	.01
Text Availability x Prior Knowledge	<1	.01
Question Delay x Prior Knowledge	<1	.02

Note. All independent variables have one degree of freedom.

Ry *s* = .430.

Between subject *df* error = 136, R^2 = .170.

Within subject *df* error = 282, R^2 = .167.

** *p* < .01

**** *p* < .001

Table 2

Partitioning of Between Subject Reading Comprehension Variance Showing the Proportion of Variance Associated with Prior Knowledge

Variable	F	Increment in Percentage of Variance Explained
	Corn	
IQ	15.45****	8.78
Prior Knowledge	21.56****	12.25
TOTAL		21.03
	City	
IQ	2.88	1.92
Prior Knowledge	7.88**	5.26
TOTAL		7.18
	Civil War	
IQ	21.78****	11.26
Prior Knowledge	32.68****	16.89
TOTAL		28.15

Note. dfe = 139.

**p < .01.

***p < .005.

****p < .001.

Table 3

Partitioning of Between Subject Reading Comprehension Variance With Significance Tests for Each Passage. Prior Knowledge and Population Group (Rural/Urban) Are Entered into the Regression in Both Orders to Show Population Bias and Its Removal

Variable	Order of Entry of Independent Variables into the Regression	F	Increment in Percentage of Variance Explained
	Corn[a]		
Rural/Urban	1	6.26*	3.69
Prior Knowledge	2	24.49****	14.43
Prior Knowledge	1	29.57****	17.42
Rural/Urban	2	1.18	.69
	City[b]		
Rural/Urban	1	12.30***	8.03
Prior Knowledge	2	<1	1.26
Prior Knowledge	1	10.38**	6.75
Rural/Urban	2	3.89	2.54
	Civil War[c]		
Rural/Urban	1	5.22*	2.87
Prior Knowledge	2	37.53****	23.52
Prior Knowledge	1	41.78****	22.99
Rural/Urban	2	<1	.53
	Civil War[d]		
Male/Female	1	3.71	2.02
Prior Knowledge	2	40.89****	22.27
Prior Knowledge	1	42.21****	22.99
Male/Female	2	2.38	1.30

[a] R^2 = .181, *df* error = 139, $\bar{X}$ = 6.94, SD = 2.73.

[b] R^2 = .093, *df* error = 139, $\bar{X}$ = 5.52, SD = 2.39.

[c] R^2 = .235, *df* error = 139, $\bar{X}$ = 7.39, SD = 3.38.

[d] R^2 = .243, *df* error = 139.

*p < .05.

**p < .01.

***p < .005.

****p < .001.

Table 4

Summary of Regression Analyses Demonstrating the Removal of Bias by the Randomization Method (summing raw scores across content areas) and the Prior Knowledge Method (partialing out the influence of prior knowledge before summing across content areas)

	Randomization Method[a]		
	F	Variance Due to Predictor	Total Variance
IQ	24.12****	5.18	35.40
Population	2.89	.62	
	Prior Knowledge Method[b]		
IQ	5.92*	.25	6.15
Population	2.19	.09	

Note. *df* error = 138.

All independent variables have one degree of freedom.

[a]Dependent variable = sum of three content scores.

[b]Dependent variable = sum of three residual content scores after the effects of prior knowledge have been removed from each.

*$p < .05$.

****$p < .001$.

Table 5

Mean[a] Correlations Between Increasingly General Vocabulary Tests and I.Q. and Reading Comprehension

Vocabulary Test	Reading Comprehension	I.Q.
Content relevant vocabulary questions (11 questions)	.39	.25
Vocabulary questions not relevant to the passage content (22 questions)	.33 (.22)	.32 (.30)[b]
All vocabulary questions (33 questions)	.35 (.31)	.37 (.32)[c]

[a]mean of the 3 correlations between vocabulary and reading comprehension scores by content area.

[b]mean correlation with 11 item vocabulary test in which the 11 items were a random selection of half of the 22 item test in order to equate reliability with the content relevant test.

[c]mean correlation with 11 item vocabulary test in which the 11 items were a random selection of one third of the total 33 items in order to equate reliability with the content relevant test.

Table 6

Vocabulary Tests as Measures of General Verbal Ability and as Measures of Prior Knowledge

Variable	F	Increment in Percentage of Variance Explained
	Corn	
IQ	16.02****	8.78
General vocabulary	13.54****	7.42
Prior knowledge	15.87****	8.70
	City	
IQ	2.83	1.92
General vocabulary	<1	.04
Prior knowledge	7.75**	5.25
	Civil War	
IQ	21.66****	11.26
General vocabulary	8.43**	4.38
Prior knowledge	25.32****	13.16

Note. General verbal ability test = score on 22 complementary vocabulary items.

Prior knowledge test = score on 11 content specific vocabulary items.

Mean percentage of variance accounted for by general vocabulary = 3.9%.

Mean percentage of variance accounted for by prior knowledge = 9%.

All independent variables have one degree of freedom.

**$p < .01$.

***$p < .005$.

****$p < .001$.

Table 7

Partitioning of Variance of Reading Comprehension

Question Type Subscores

Variable	F	Increment in Percentage of Variance Explained
	Within	
Passage Contrast 1	29.38****	1.10
Passage Contrast 2	33.95****	1.27
Prior knowledge	16.56****	.62
Scriptal vs. Textual questions (Q1)	105.92****	3.95
Text explicit vs. Text implicit (Q2)	31.99****	1.19
Centrality	<1	.03
Prior knowledge x Q1	2.55	.10
Prior knowledge x Q2	<1	.02
Prior knowledge x Centrality	7.33**	.27
Q1 x Centrality	4.59*	.17
Q2 x Centrality	3.68	.14
Prior knowledge x Centrality x Q1	<1	.03
Prior knowledge x Centrality x Q2	<1	--
Prior knowledge x Centrality x Text availability	<1	.01
Prior knowledge x Centrality x Question delay	<1	--
Q1 x Text availability	16.10****	.60
Q1 x Question delay	<1	.01
Q2 x Text availability	1.10	.04
Q2 x Question delay	2.28	.09
Centrality x Text availability	7.62**	.28
Centrality x Question delay	2.63	.10

Table 7 (continued)

Variable	F	Increment in Percentage of Variance Explained
Prior knowledge x Centrality x Text availability	9.85***	.37
Prior knowledge x Centrality x Question delay	<1	.03
Prior knowledge x Q1 x Text availability	<1	.02
Prior knowledge x Q1 x Question delay	2.95	.11
Prior knowledge x Q2 x Text availability	<1	.02
Prior knowledge x Q2 x Question delay	2.09	.08
Q1 x Centrality x Text availability	8.32**	.31
Q1 x Centrality x Question delay	<1	.03
Q2 x Centrality x Text availability	<1	.04
Q2 x Centrality x Question delay	<1	.01

Note. All independent variables have one degree of freedom.

Ry s = .0365.

Between subjects R^2 = .082, df error = 150.

Within subjects R^2 = .1052, df error = 2,388.

*p < .05.

**p < .01.

***p < .005.

****p < .001.

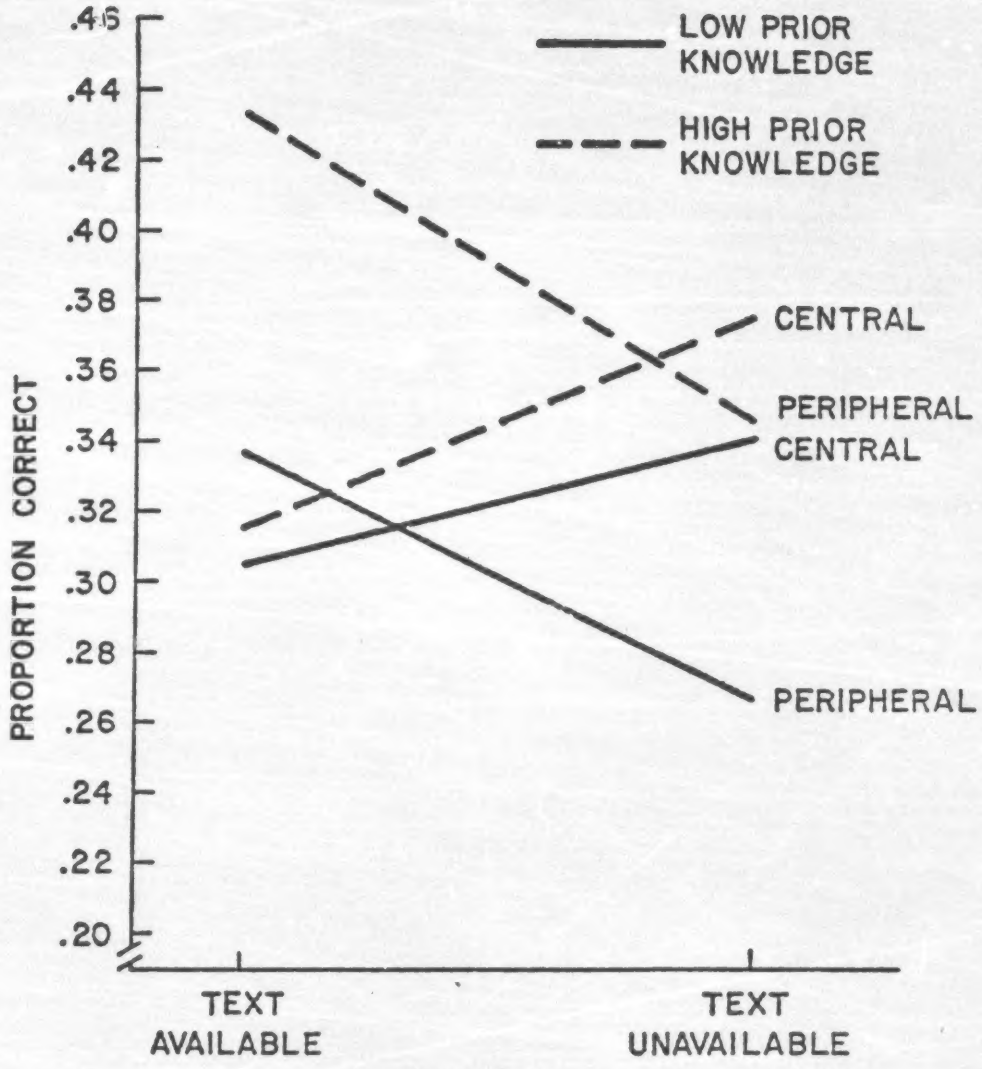

Figure 1. The three-way interaction between reader prior knowledge, question centrality and long-term memory demands of the task on proportion of questions correct.

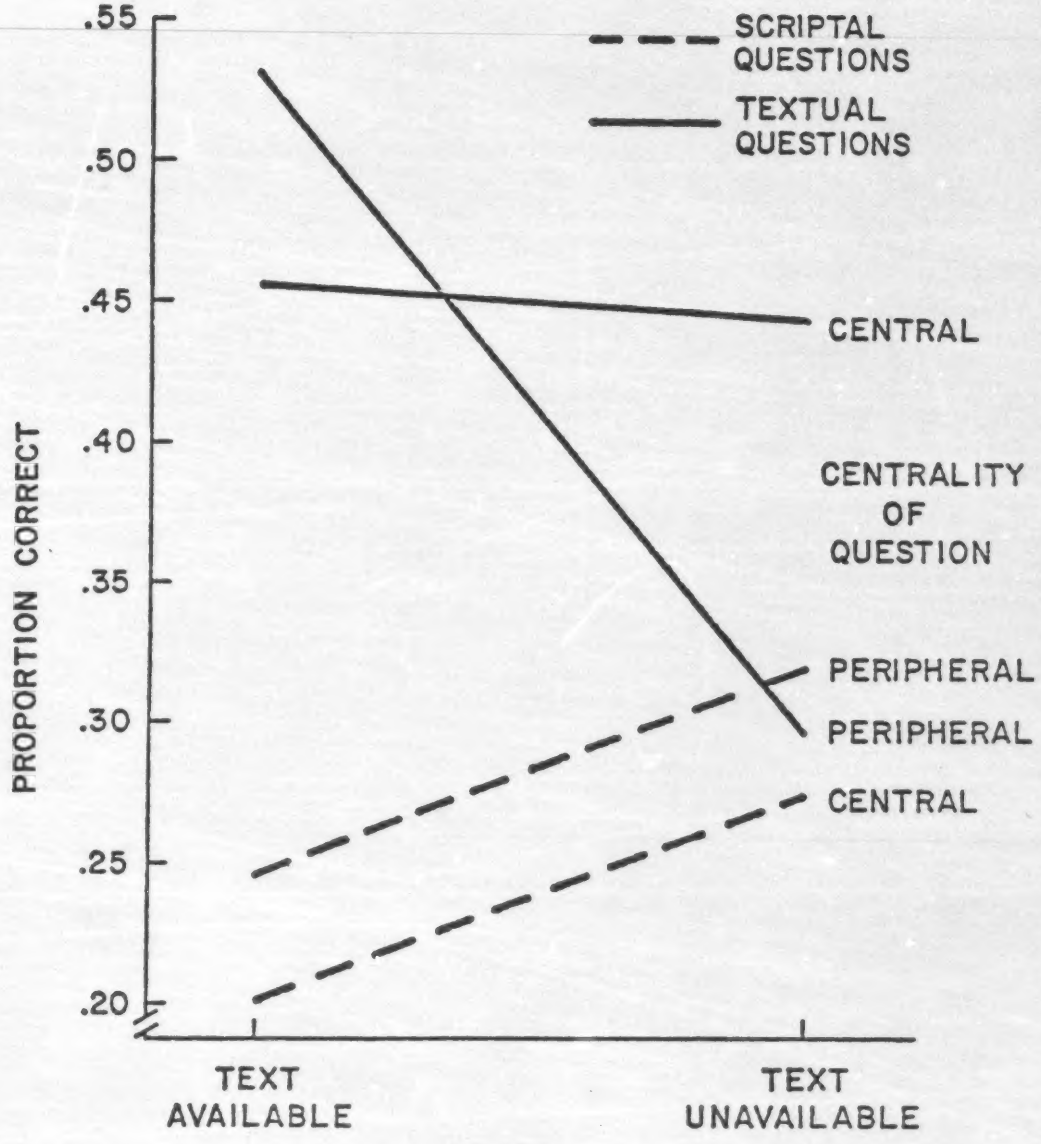

Figure 2. The effects of the interaction between question type, centrality of the question, and the long-term memory demands of the task, on proportion of questions correct.